Assessment of the Inter-Rater Reliability of the Behavioural Assessment and Treatment Taxonomy (BATT)

Ateeq Awan. B. A. Hons.

Centre for Applied Disability Studies

Submitted in partial fulfilment
of the requirements for the degree of

Master of Arts

Faculty of Social Sciences, Brock University
St. Catharines, Ontario

Abstract

Applied Behaviour Analysis (ABA) is an evidence-based approach for treating a range of behavioural challenges and skill deficits commonly associated with developmental disabilities such as Autism Spectrum Disorder (ASD). ABA-based interventions are typically tailored to the socially significant needs of the individual receiving treatment. This individualized nature of treatment delivery can make it challenging for wide-scale program evaluation within and across different ABA service providers. The Behavioural Assessment and Treatment Taxonomy (BATT) is a tool designed specifically for the purpose of standardizing the documentation of assessment and treatment strategies and enabling systematic evaluation of treatment outcomes. The purpose of this study was to assess the inter-rater reliability of the BATT using a retrospective study design (N = 112). The results demonstrated acceptable level of inter-rater reliability (κ > .80) across a number of target variables, although additional refinement of the BATT coding scheme is still required.

Acknowledgements

First of all, I would like to thank my thesis supervisor Dr. Rosemary Condillac for her endless support and guidance throughout this journey. I would also like to express my appreciation to the various faculty members at Brock University who helped make my graduate school experience most memorable and rewarding. Additionally, I would like to thank the participating agency that provided the opportunity for me to conduct my research and write this thesis. Lastly, I would like to express my sincerest gratitude to my family for their love and support and for giving me the strength to reach this point in my life.

Table of Contents

List of Tables

Assessment of the Inter-Rater Reliability of the

Behavioural Assessment and Treatment Taxonomy (BATT)

Autism spectrum disorder (ASD) is commonly associated with a number of challenging behaviours including aggression and self-injury (Heyvaert, Saenen, Campbell, Maes, & Onghena, 2014). Families of individuals with ASD who exhibit dangerous behaviours such as aggression have a higher likelihood of experiencing financial challenges and increased levels of stress (Lecavalier, Leone, & Wiltz, 2006). Although there are a number of approaches for treating challenging behaviours in individuals with ASD, the most common treatments include psychotropic and behavioural interventions (Matson & Jang, 2014). There is mixed evidence for the efficacy of commonly prescribed medications for treating aggression in individuals with ASD and their use requires careful consideration given the number of associated side effects (Nevels, Dehon, Alexander, & Gontkovsky, 2010; Parikh, Kolevzon, & Hollander, 2008). Conversely, there is substantial evidence for the effectiveness of behavioural interventions based on principles of Applied Behaviour Analysis (ABA) for treating challenging behaviours in individuals with ASD (e.g., Brosnan & Healy, 2011; Sturmey, 2012).

ABA-based interventions typically begin with functional assessment procedures aimed at identifying environmental events maintaining an individual's challenging behaviours as well as determining appropriate replacement behaviours and skills. ABA practitioners in community-based ABA programs work closely with parents and care-providers of individuals with ASD to design and implement interventions for socially significant challenges (Feldman, Condillac, Tough, Hunt, & Griffiths, 2002; Kaale, Smith, & Sponheim, 2012; Sheinkopf & Siegel, 1998). The results of the functional assessment process allow ABA practitioners to individualize

treatments to the specific needs of the individual with ASD while taking into consideration the capacity of the care-providers for treatment implementation.

ABA-based services are typically delivered using two different models: (1) clinicians delivering treatment directly to the individual with challenging behaviours (i.e., direct service model) or (2) the clinician providing training to the care-providers on necessary skills needed to deliver the intervention (i.e., indirect or mediator model; Baker, Leff, Bevans, & Power, 2014). Although several studies have demonstrated the effectiveness of interventions based on the mediator model (Feldman et al., 2002; Jocelyn, Casiro, Beattie, Bow, & Kneisz, 1998), there has been limited attention paid to the critical components of mediator implemented interventions including the competence of the ABA practitioner and standardization of behavioural consultation practices (Kratochwill & Van Someren, 1995; Sanetti & Kratochwill, 2008; Schulte, 2007). Furthermore, data on treatment integrity are generally lacking (McIntyre, Gresham, DiGennaro, & Reed, 2007; Wheeler, Baggett, Fox, & Blevins, 2006) and the need to document and report treatment integrity data is a crucial aspect of evaluating community-based ABA programs for individuals with ASD.

One of the requirements of wide-scale evaluation of ABA programs is consistency in documentation standards and practices employed by different ABA service providers. Although there has been a recent movement in health psychology and behavioural medicine towards improving documentation procedures (Michie et al., 2013), there is still a need for a more comprehensive measure that can allow consistent monitoring and evaluation of behavioural treatments within and across service providers. To meet this need, a system of sustainable program evaluation tools was developed (Condillac, 2009a). This system includes three tools that can be used together as a comprehensive program evaluation strategy. The Target Behaviour

Severity Scale (TBSS; Condillac, 2009d) is a measure of the severity of problem behaviour. The Impact on Quality of Life (IQOL; Condillac, 2009b) measures the impact of problem behaviour on quality of life of the individual and those in their natural environments. The Behavioural Assessment and Treatment Taxonomy (BATT; Condillac, 2009c) is a tracking/supervision tool for behavioural assessment, treatment, monitoring, and implementation strategies, and outcomes. The purpose of the current study is to evaluate the inter-rater reliability of the BATT using a retrospective study design[1].

Literature Review

The aim of this section is to provide an overview of research highlighting the need for systematic and wide-scale program evaluation of ABA-based behavioural services provided to individuals with ASD. First, a description of the nature of developmental disabilities (including ASD) will be provided along with a depiction of the associated behavioural challenges and skill deficits. Second, research on psychopharmacological and behavioural interventions used to treat behavioural challenges in individuals with ASD will be presented. Third, the evidence base for behavioural interventions based on the principles of ABA will be described. Fourth, the service delivery models for behavioural interventions will be discussed and evidence for the efficacy of the mediator model will be provided. Fifth, research on the evaluation of mediator-implemented behavioural interventions will be presented and gaps in existing research will be identified. Lastly, the BATT will be offered as a potential tool for addressing the existing barriers to wide-scale program evaluation.

[1] This study is being conducted within a larger study by Condillac (2015) focused on the establishment of reliability and validity of these program evaluation measures.

Developmental Disability, Intellectual Disability, and Autism Spectrum Disorder

Developmental disability (DD) is an umbrella term that encompasses a wide range of neurodevelopmental disorders such as autism spectrum disorder (ASD) and intellectual disability (intellectual developmental disorder; American Psychiatric Association [APA], 2013). The onset of these disorders typically occurs during early development and may include deficits in social, personal, academic, and occupational domains (APA, 2013). The deficits can vary from specific learning challenges in executive functioning to more global impairments in intelligence (APA, 2013).

Individuals diagnosed with intellectual disability generally exhibit early onset of impairments in (1) cognitive functioning such as reasoning, problem solving, and abstract thinking; and (2) adaptive functioning which may result in challenges with developing personal independence and skills required for activities of daily living (e.g., communication, social integration, etc.; APA 2013).

ASD is a neurodevelopmental condition that is characterized by: (1) deficits in social communication and interaction skills, and (2) restricted behaviour, interests, and activities (APA, 2013). Social and communication difficulties can include impairments in: (a) social-emotional reciprocity (e.g., abnormal social approach or reduced affect); (b) nonverbal behaviours related to social communication (e.g., atypical eye contact and facial expressions); and (c) development and maintenance of social relationships (e.g., difficulty making friends or lack of desire to engage in social interaction; APA, 2013). Furthermore, restricted and repetitive behaviours include: (a) stereotyped or repetitive motor movements including use of speech and other objects (e.g., echolalia, motor stereotypy); (b) inflexibility with routines and insistence on sameness (e.g., extreme distress due to small changes in routine); (c) restricted and fixated interests that are

beyond normal level of intensity (e.g., preoccupation with specific objects); and (d) hyper- or hypo-reactivity to sensory stimuli (e.g., elevated tolerance for pain or hyper-responsiveness to auditory stimuli; APA, 2013). Lastly, other criteria required for a diagnosis include: (a) existence of the symptoms mentioned above at an early developmental age that were not a product of intellectual developmental disorder or global developmental delay (although intellectual disability may co-occur with ASD), and (b) clinically significant impairment in social functioning (APA, 2013).

ASD and Associated Challenging Behaviours in Children and Adults

Children, youth, and adults with ASD are considered to have a lifelong condition (Tager-Flusberg, 2014) that is commonly associated with a number of challenging behaviours including aggression, stereotypy, and self-injury (Heyvaert et al., 2014). Studies investigating prevalence of aggression in ASD have found high rates of challenging behaviours associated with the disorder. For example, in a sample of 1380 children and adolescents with ASD, 56% were reported by parents to be currently engaging in aggression towards caregivers, while 68% were reported to have engaged in aggression in the previously (Kanne & Mazurek, 2011). Similarly, in a sample of 1534 children and adolescents with ASD, Mazurek, Kanne, and Wodka (2013) found aggressive behaviour reported for 47% of individuals aged 14-17 years. In addition, individuals demonstrating aggressive behaviour also experienced greater challenges with social and communication skills and exhibited sensory and gastrointestinal problems, in comparison to individuals who did not engage in aggression (Mazurek et al., 2013). This pattern of challenging behaviours and associated difficulties has also been demonstrated in adults with ASD, with aggression posing a greater concern in adulthood due to longer history of exhibiting the behaviour (Matson & Jang, 2014).

In addition, challenging behaviours such as aggression can pose a significant threat to the welfare of individuals with ASD and their families (Kanne & Mazurek, 2011; Mazurek et al., 2013). A number of domains related to the experiences of families with a child with DD such as ASD have been evaluated regarding physical health and emotional well-being, community involvement and support, parental stress, and financial well-being (Baker et al., 2003; Brown, MacAdam-Crisp, Wang, & Iarocci, 2006; Petry, Maes, & Vlaskamp, 2005). Research in these areas has shown that families of children with ASD are more likely to experience greater financial expenses, restrictions in social and community involvement, and increased levels of stress (Lecavalier et al., 2006).

The relationship between family quality of life and challenging behaviours is one that is complex and still poorly understood (Baker et al., 2003; Sikora et al., 2013; Zaidman-Zait et al., 2014). In a sample of 293 children with ASD, Baker et al. (2003) discovered that high levels of parental stress were associated with an increase in challenging behaviours over time and high rates of challenging behaviours were associated with an increase in parental stress.

More recently, researchers investigating the impact of challenging behaviours on family functioning have focused on the relationship between parental stress and the differential impact of internalizing behaviour problems (e.g., withdrawal, anxiety) and externalizing behaviour problems (e.g., aggression, self-injury) in children with ASD (Sikora et al., 2013; Zaidman-Zait et al., 2014). For example, Zaidman-Zait et al. (2014) found a modest bidirectional relationship between parenting distress (i.e., stress directly related to parenting of the child with ASD) and both externalizing and internalizing behaviour problems from 12-months after diagnosis until the age of 6-years. In addition, when considering the impact of intensity of behaviour problems on parental stress and quality of life, Sikora et al. (2013) found significant group differences for low

versus high externalizing behaviours: higher ratings of externalizing behaviour problems on the Child Behaviour Checklist were found to be associated with greater impact on family quality of life (based on ratings on the Family Impact Questionnaire-Revised; FIQ-R).

Overall, the studies reviewed above provide strong evidence for the association between challenging behaviours and level of parental stress. Hence, early intervention and treatment of challenging behaviours is more likely to lead to an improvement in the quality of life of individuals with DD and their families (Feldman et al., 2002).

Treatment of Challenging Behaviours and Skill Deficits in Individuals with ASD

Psychotropic and behavioural interventions represent the most common treatment options for reducing challenging behaviours such as aggression in children and adults with developmental disabilities such as ASD (Matson & Jang, 2014). Psychotropic interventions involve prescription of various medications for the presenting behaviour problems. In order to assess the effectiveness of a particular medication on changes in behaviour, clinicians may employ a standardized measure such as the aberrant behavior checklist (ABC; Aman, Singh, Stewart, & Field, 1985). The ABC consists of five subscales including: (I) irritability, agitation, crying; (II) lethargy, social withdrawal; (III) stereotypic behavior; (IV) hyperactivity, noncompliance; and (V) inappropriate speech (Aman et al., 1985). Changes in aggressive behaviour are generally detected using scores on Scale I (i.e., irritability).

Clinical recommendations for physicians on use of psychotropic interventions include regular reviews of the effectiveness of the medications as prolonged usage in some adults with DD can increase risk of harm to the individual (Sullivan et al., 2011). Furthermore, given the challenges associated with diagnosing psychotic disorders in individuals with DD, Sullivan et al. (2011) recommend against the use of antipsychotic medication for treating challenging

behaviours in the absence of a confirmed diagnosis. Other studies in this area have shown that there is limited evidence for the use of many psychotropic medications. For example, Parikh et al. (2008) conducted a review of 21 randomized placebo-controlled trials for commonly prescribed medications (such as risperidone and methylphenidate) for treating aggression in individuals with ASD. The authors found mixed results for the efficacy of several medications while noting various associated side effects (e.g., weight-gain; Parikh et al., 2008). Taken together, these studies highlight the importance of giving careful consideration when using psychotropic medications for treating behavioural challenges in individuals with developmental disabilities (Parikh et al., 2008; Sullivan et al., 2011).

Behavioural interventions, on the other hand, represent an effective treatment option for reducing challenging behaviours and teaching replacement skills (Sturmey, 2012). For example, in a review of various types of interventions for psychopathology in individuals with developmental disabilities, Sturmey (2012) found strong evidence for the effectiveness of behavioural interventions and limited evidence for use of other treatments such as cognitive therapy and psychotropic medications. In addition, although there could be beneficial effects (i.e., additive effects) of combining psychotropic and behavioural treatments (Weeden, Ehrhardt, & Poling, 2009), research in this area is currently lacking (Courtemanche, Schroeder, & Sheldon, 2011).

Several studies have provided evidence for the effectiveness of behavioural interventions for treating challenging behaviours. In a review of 18 behavioural intervention studies, Brosnan and Healy (2011) grouped behavioural interventions into three main categories: (1) antecedent manipulations (e.g., choice making), (2) reinforcement based strategies (e.g., differential reinforcement), and (3) consequential control (e.g., extinction; Brosnan & Healy, 2011). Overall,

behavioural interventions were found to be effective in reducing aggression across all reviewed studies. However, generalizability of results beyond treatment were limited as only a few studies reported follow-up data on rates of challenging behaviours (Brosnan & Healy, 2011). In a more recent review of 213 behavioural intervention studies, Heyvaert et al. (2014) found behavioural interventions on average to be effective in achieving behaviour change regardless of the type of behaviour targeted. Therefore, the studies reviewed above provide evidence for the efficacy of behavioural interventions for treating challenging behaviours in individuals with developmental disabilities.

ABA-based Behavioural Interventions for Children and Adults with ASD

There is substantial research demonstrating the effectiveness of behavioural interventions based on the principles of Applied Behaviour Analysis (ABA; Sturmey, 2012). ABA is an evidence-based scientific approach for treating socially significant behavioural challenges (Baer, Wolf, & Risley, 1987). ABA-based behavioural interventions begin with an assessment of environmental contingencies impacting and maintaining behaviour. ABA assessment procedures involve a number of different approaches including indirect measures such as questionnaires, direct observation of the behaviour, or experimental manipulation of contingencies thought to maintain behaviour (Cooper, Heron, & Heward, 2012).

Indirect assessments generally comprise of checklists or questionnaires such as the Questions About Behavioural Function (QABF; Matson & Vollmer, 1995), the Motivation Assessment Scale (MAS; Durand & Crimmins, 1988), or structured interviews such as the Functional Assessment Interview (Hanley, 2012). Although indirect measures are the simplest to conduct and can provide useful contextual information (Hanley, 2012), there is limited validity

of such assessments for determining the function of challenging behavior (in comparison to experimental methods; Wightman et al., 2014).

Another form of assessment involves the use of descriptive analyses or direct observation of the behaviour in the natural environment (Cooper et al., 2007). Most common use of descriptive assessments involves observing the events that precede (i.e., antecedents) and follow (i.e., consequences) the behaviour (Cooper et al., 2007). Although descriptive analyses may be preferred or used in conjunction with indirect assessments, the validity of their use for assessing behavioural function is also limited. For example, in a direct comparison of descriptive and experimental analyses, Thompson and Iwata (2007) found that descriptive analyses only correctly predicted the consequences maintaining challenging behaviours in 25% of the cases. Therefore, caution should be exercised when considering use of descriptive (and/or indirect) assessments for determining the function of problem behaviour.

The use of experimental manipulation for assessing the function of challenging behaviours is known as functional analysis (FA; Hanley, Iwata, & McCord, 2003). An FA involves the experimental manipulation of one or more environmental contingencies that are thought to maintain the behaviour of interest (Iwata, Dorsey, Slifer, Bauman, & Richman, 1982/1994; Neidert, Dozier, Iwata, & Hafen, 2010). In this manner, an FA is designed to briefly produce conditions that result in reinforcement of the challenging behaviour (Hanley, 2012). For dangerous behaviours such as aggression and self-injury, this can pose ethical concerns due to the risk of harm to individual participating in the FA (Hanley, 2012). In response to such concerns, a number of variations have been developed such as the use of FAs designed to evoke precursor behaviours (i.e., behaviours at the low end of escalation; Najdowski, Wallace, Ellsworth, MacAleese, & Cleveland, 2008) and latency-based functional analyses (e.g.,

Thomason-Sassi, Iwata, Neidert, & Roscoe, 2011), where an FA session is terminated upon the first occurrence of the target behaviour. These variations may reduce the risk of harm and associated ethical concerns for more severe forms of dangerous behaviours (Hanley, 2012).

A number of studies have demonstrated the benefits of using FAs for assessing challenging behaviours (e.g., Hanley et al., 2003; Harvey, Boer, Meyer, & Evans, 2009; Heyvaert et al., 2014). For instance, Hanley et al. (2013) conducted a review of 277 behavioural intervention studies involving the use of an FA for assessing the function of challenging behaviours (with 40% of the studies identifying aggression as the target). Approximately 96% of studies were found to report successful assessment of function through the use of an FA (Hanley et al., 2013). Similarly, in a more recent review of 435 behavioural intervention studies, Beavers, Iwata, and Lerman (2013) found that behavioural function was successfully identified in approximately 92% of studies utilizing an FA.

Although experimental manipulation of environmental contingencies through an FA have been shown to be highly effective for assessing behavioural function (Heyvaert et al., 2014), clinicians must consider: (1) the training and resources needed to conduct the FA, (2) the dangerousness of the behaviour (e.g., aggression, self-injury), and (3) the ability to create a controlled environment for systematically manipulating conditions (Wightman, Julio, & Virués-Ortega, 2014). Overall, the studies reviewed highlight the necessity of using function-based treatment methods when targeting challenging behaviours in individuals with ASD.

Service Delivery Models for Behavioural Interventions

Community-based behavioural intervention services are typically delivered to individuals with ASD using two basic models. The *direct* service model involves a therapist delivering the treatment directly to the client on an individual basis or within a group setting while the *indirect*

or *mediator* model involves the therapist providing training and consultation to mediators (e.g., parents, teachers, care providers) who in turn act as the primary interventionists (Baker et al., 2014).

There is considerable research demonstrating the efficacy of including mediators in various aspects of behavioural interventions such as: directing the intensity of treatment (Sallows & Graupner, 2005), participating in knowledge and/or performance based training (Eikeseth, Smith, Jahr, & Eldevik, 2002; Ingersoll & Dvortcsak, 2006; Lovaas, 1987), and directly implementing the interventions under professional supervision (Feldman et al., 2002; Kaale et al., 2012; Sheinkopf & Siegel, 1998). Several studies have shown that mediator-implemented interventions in the natural setting (e.g., home, school) can be successful in decreasing challenging behaviours and increasing prosocial skills in individuals with ASD (Feldman et al., 2002; Jocelyn et al., 1998). A review by Reichow, Servili, Yasamy, Barbui, and Saxena (2013) of 13 behavioural intervention studies involving treatments implemented by mediators showed significant effects related to improvements in adaptive skills and development (e.g., IQ, language, developmental progress) of children with severe cognitive delays.

In their meta-analysis of 109 published studies with 230 participants, Carr et al. (1999) reviewed a number of different factors that can impact the ecological validity of behavioural treatments including treatment setting, context, and delivery agents. Although significant differences were not found based on setting (i.e., typical settings such as home or school versus atypical settings such as hospitals and psychiatric wards) or context (i.e., treatment applied in some versus all relevant contexts pertaining to the target behaviours), interventions delivered by typical agents (e.g., parents, teachers, direct care staff) were found to have a higher success rate

(61%) in comparison to interventions delivered by atypical agents (e.g., psychologists, behaviour specialists; 44.3% success).

Mediator involvement through training on behavioural principles and strategies can also impact the effectiveness and generalizability of interventions and lead to other positive outcomes (Strauss, Mancini, Fava, & SPC Group, 2013). For example, a recent review of 21 early intensive behavioural intervention (EIBI) studies revealed a larger overall effect size for IQ and adaptive behaviours for treatments that included parent-training for the purpose of generalization and maintenance of the intervention at home (Strauss et al., 2013). Furthermore, a review by McConachie and Diggle (2007) of 12 behavioural intervention studies found that inclusion of parent training led to gains in child social communication skills, parental knowledge of ASD, improved parent-child interactions, and reductions in maternal depression. Hence, mediator involvement may not only increase treatment efficacy but also lead to improvements in family quality of life for individuals with ASD (Feldman et al., 2002).

Overall, mediator-implemented behavioural interventions represent an ecologically and socially valid form of treatment (Carr et al., 1999). Given the limited resources in healthcare funding and high costs of treatments implemented by behavioural professionals (i.e., direct service model), the mediator model represents the treatment delivery of choice for achieving generalized reduction in challenging behaviours and skill deficits in individuals with ASD. However, given the range of skills and training required by mediators, the success of these interventions should be monitored through ongoing program evaluation.

Evaluating Mediator-Implemented Interventions

Evaluating community-based behavioural interventions that are based on the mediator model requires consideration of two distinct elements: (1) the competency of the clinician in

adequately (a) assessing the target behaviour to determine function, (b) designing an appropriate intervention plan, and (c) training the mediators on the intervention procedures; and (2) the integrity with which the mediators adopt the recommendations of the consultant and deliver the intervention to the client (Baker et al., 2014). This is in contrast to the direct service model which only requires assessment of the performance of the therapist in relation to treatment delivery. Hence, mediator based interventions present a greater level of complexity when attempting to identify the active ingredients responsible for the success (or failure) of such treatments.

Although outcome studies have shown that interventions based on the mediator model can be effective in teaching new skills and targeting challenging behaviours in individuals with ASD (Carr et al., 1999; Feldman et al., 2002), a number of reviews have highlighted the need for measuring and reporting treatment integrity (e.g., McIntyre et al., 2007; Wheeler et al., 2006). Studies evaluating intervention outcomes have consistently demonstrated that higher levels of treatment integrity are associated with improved outcomes (Biggs, Vernberg, Twemlow, Fonagy, & Dill, 2008; DiGennaro, Martens, & Kleinmann, 2007; Wilder, Atwell, & Wine, 2006). For example, a study by Wilder et al. (2006) evaluated the impact of varying levels of treatment integrity (i.e., 100%, 50%, and 0%) on compliance with requests in two typically developing children. The results revealed that compliance was directly related to the level of treatment integrity with highest rates of compliance (mean of 91% and 78% respectively for the two participants) found when the procedures were implemented with 100% integrity.

Research on treatment integrity has generally focused on two types of integrity failures: errors of omission and errors of commission (Fryling, Wallace, & Yassine, 2012). For mediator-implemented interventions, errors of omission refer to the failure in administering specific components of a treatment program (e.g., missing opportunities to reinforce the target

behaviour), while errors of commission refer to the incorrect application of the treatment plan (e.g., reinforcing challenging behaviours; Fryling et al., 2012; St. Peter Pipkin, Vollmer, & Sloman, 2010).

The impact of treatment integrity failures can vary depending on the type of errors involved. Northup, Fisher, Kahng, Harrell, and Kurtz (1997) investigated the role of errors of omission on reinforcement and punishment procedures. Northup et al. (1997) systematically varied the schedules of reinforcement and punishment while implementing each procedure for 100%, 50%, and 25% of the recommended treatment intensity. The results indicated that only at low levels of integrity (i.e., 25%) were there any detrimental effects of errors of omission (Northup et al., 1997). Similarly, DiGennaro Reed, Reed, Baez, and Maguire (2011) investigated the impact of varying levels of errors of commission (i.e., 100%, 50%, or 0% errors of commission) on error correction procedures during a discrete-trial training program. The authors found that performance improved for all three participants when the procedure was implemented without any integrity failures (i.e., 0% errors of commission) while performance was equally low during both 50% and 100% conditions, indicating that 50% errors of commission were just as harmful as carrying out the procedure without 0% integrity (i.e., 100% errors of commission).

In a study comparing the impact of errors of omission and commission on treatment outcomes, St. Peter Pipkin et al. (2010) applied different treatment integrity failures to a differential reinforcement procedure implemented in experimental (i.e., laboratory) and non-experimental (i.e., clinical) settings. Errors of omission were defined as the failure to deliver reinforcers based on treatment schedule while commission errors involved reinforcing challenging behaviours. The authors found that in both settings, reinforcing challenging behaviours (i.e., errors of commission) was found to be more problematic than failing to provide

reinforcement for appropriate behaviours (i.e., error of omission). Furthermore, the order in which the two conditions were implemented influenced the results such that the detrimental effects of the low integrity condition were reduced when it was preceded by the condition with high integrity.

Overall, treatment integrity research has generally highlighted the need to document and report treatment integrity data (McIntyre et al., 2007; Wheeler et al., 2006) as high levels of integrity in treatment implementation are associated with improved outcomes (Biggs et al., 2008; DiGennaro et al., 2007; Wilder et al., 2006). Furthermore, depending on the type of integrity failures involved (i.e., omission or commission errors), treatment integrity issues can differentially impact treatment outcomes (DiGennaro et al., 2011; Northup et al., 1997; St. Peter Pipkin et al., 2010). Hence, documenting treatment integrity data and analysing the types of errors should be considered for improving mediator implemented treatments.

Research Gaps in the Evaluation of Mediator Implemented Interventions

Although there has been a recent shift in outcome research towards the assessment of treatment integrity in behavioural interventions (DiGennaro et al., 2011; Northup et al., 1997; St. Peter Pipkin et al., 2010), studies in this area have generally focused on mediator adherence to treatment procedures as an important variable impacting treatment outcomes (McIntyre et al., 2007; Wheeler et al., 2006). However, several reviews have identified the need to provide greater consideration to other critical components of the behavioural assessment and treatment process such as therapist competence and standardization of behavioural consultation practices (Kratochwill & Van Someren, 1995; Sanetti & Kratochwill, 2008; Schulte, 2007). This applies critically to mediator implemented treatments as the effectiveness of the intervention can depend on a number of therapist competencies such as: (1) adequate assessment and identification of

target behaviours, (2) selection of appropriate evidence-based intervention strategies; (3) training mediators on intervention procedures; (4) monitoring intervention progress by measuring changes in target behaviour; and (5) evaluating mediator adherence to treatment recommendations (Condillac, 2009d).

In addition to therapist, mediator, and client variables, the validity of functional behaviour assessment procedures depends on whether different therapists can provide similar conclusions regarding the functional relationships identified during assessment (Shriver, Anderson, & Procter, 2001). Shriver et al. (2001) discuss several characteristics of the assessment process that can contribute to the validity of behavioural interventions. Firstly, the types of assessment strategies utilized should enable practitioners to operationally define target behaviours and select appropriate replacement behaviours. Secondly, the measures used in the assessment process should cover a variety of environmental stimuli that may be important for determining functional relations. Thirdly, the types of environmental events included in the analysis should be representative of the setting where the intervention is to be implemented. Overall, the goal of the behavioural assessment and consultation process is to identify functional relationships between target behaviours and environmental stimuli and to utilize this information for implementing the appropriate intervention strategies.

Another aspect relevant to the validity of the functional behaviour assessment procedures is the relationship between assessment and treatment outcomes (i.e., treatment validity; Shriver et al., 2001). Treatment validity refers to the degree to which assessment methods positively impact the outcome of an intervention (Hayes, Nelson, & Jarrett, 1987). Hayes et al. (1987) describe a number of benefits of evaluating the treatment validity of functional behaviour assessment procedures. Firstly, treatment validity data can inform decisions regarding selection

of various assessment methods for determining functions of target behaviours. Research in this area has typically demonstrated the efficacy of experimental functional analyses over the use of indirect or descriptive procedures (Beavers et al., 2013; Hanley et al., 2003). Secondly, treatment validity data can inform the efficacy of various treatment methods for specific target behaviours. To this point, there is considerable research on effectiveness of specific behavioural intervention strategies (e.g., functional communication training; FCT; Carr & Durand, 1985) for various target behaviours based on different functions of behaviour. Thirdly, treatment validity data may contribute to assessment of social validity of behavioural interventions. Although research in this area is not as extensive, greater emphasis is now being given to factors such as quality of life (Feldman et al., 2002) and treatment acceptability (Kratochwill & Van Someren, 1995).

More recently, there has been a call for accurate and detailed reporting of treatment validity data in the fields of health psychology and behavioural medicine (Michie, Francis, Hardeman, & Eccles, 2008). In order to improve documentation procedures for identifying the content of interventions that are most likely responsible for the observed behaviour change, Michie et al. (2013) developed a measure comprised of a hierarchical taxonomy of 93 behavioural (e.g., shaping, chaining) and cognitive-behavioural (e.g., cognitive dissonance, self affirmation) treatment methods. Although a hierarchical taxonomy of behaviour change strategies can be useful for developing a systematic approach towards treatment of complex challenges, there is still a need for a more comprehensive measure that can capture information regarding client characteristics, assessment methodology, intervention implementation, and monitoring procedures. This type of measurement tool would be highly valuable for wide-scale program evaluation of community-based behavioural services. Given that such a tool would be implemented by clinicians with varied educational backgrounds and experiences, a measurement

tool developed for this purpose should: (a) require minimal training, (b) involve minimal effort to complete, (c) be accurate in documenting intervention progress, (d) be sensitive to changes over time, (e) and demonstrate high level of inter-rater reliability.

Rationale

Given the individualized nature of ABA-based behavioural interventions, systematic program evaluation across service recipients can be a challenging endeavour for service agencies and regulatory bodies. Although there are considerable efficacy data available on the use of specific ABA procedures for specific problems, there is limited effectiveness data from real-world implementation largely because these data are not readily available. In response to this need, Condillac (2009c) developed the Behavioural Assessment and Treatment Taxonomy (BATT) which is a tool designed specifically for the purpose of helping clinicians to: (1) document the types of assessment and treatment strategies they use in their clinical practice; (2) monitor the progress and outcomes of their treatments; and (3) gather data that may be utilized for wide-scale program evaluation.

The BATT was originally designed for prospective use by clinicians during their practice (Condillac, 2009c). The initial development of the tool involved assessing face validity of the BATT. Behaviour consultants working in community-based agencies reviewed the BATT and provided feedback with respect to the sections and items that were included (N = 39; Condillac, 2009a). However, organizations involved in the pilot suggested that the BATT would be helpful in their retrospective program evaluation efforts. Hence, prior to this study, refinements were made to the BATT coding scheme for retrospective evaluation and to align with specific behavioural procedures. Additional agency specific items were also added to the list of assessments.

Therefore, the purpose of this study was to examine the inter-rater reliability of the BATT when applied to retrospective review of client case records. The data used in the current thesis was drawn from data collected from a larger companion thesis (Boutsis, 2016) involving program evaluation of a participating multi-site agency delivering ABA-based services to children and adolescents with ASD.

Research Questions

1. **Can the BATT be used retrospectively with a high degree of reliability to extract necessary information for program evaluation?** The first hypothesis is that the BATT can be used retrospectively to extract accurate information for file review.
2. **Can the BATT coding scheme be used with a high degree of reliability to code key variables for program evaluation?** The second hypothesis is that the BATT will yield an acceptable level of inter-rater reliability for coding and capturing information contained in the files reviewed.

Method

Study Design and Setting

The current study was approved by the research ethics boards (REB) of Brock University and a participating agency which offers a number of different ABA-based programs to individuals with developmental disabilities in two geographical locations in Ontario, Canada. The study employed a retrospective design to assess the inter-rater reliability of the BATT using client records provided by the participating agency. The data utilized in this study pertained to one of the agency's ABA-programs that is based on the mediator model and involves providing short-term behavioural intervention services (i.e., 12 weeks in duration) to children and adolescents with ASD. The program also exclusively focuses on providing services to

individuals with challenging behaviours rather than focusing on other characteristic deficits of ASD (e.g., adaptive skills, social skills, etc.). Furthermore, the data provided by the agency were also used for a companion thesis (Boutsis, 2016) that involved evaluating the behavioural outcome of the ABA-based program. All data collection occurred at the premises of the participating agency in order to maintain confidentiality and privacy of client information.

Participants

Once ethics approval was granted, a sample of 422 client case records was obtained from the participating agency and subjected to the following inclusionary criteria: (a) diagnosis of autism; (b) clients aged 17 years and younger; (c) agency service agreement signed between November 1, 2013 and July 30, 2015; (d) evidence of service provided (i.e., indication that goal was met or not met); (e) target behaviours and goals pertaining to the behaviour domain (i.e., primarily clients with behaviours to decrease); (f) files containing the following three agency specific documents: (1) Service Provision Report (SPR; i.e., document summarizing nature of services provided by the agency), (2) Functional Behaviour Assessment (FBA) document, and (3) Behaviour Intervention Plan (BIP) document; (g) minimum of three baseline and treatment data points; and (h) files with daily occurrence data.

The exclusionary criteria included: (a) referral area other than behaviour reduction; (b) files with missing documents such as the SPR, FBA, or BIP; (c) fewer than three baseline or treatment data points; and (d) files with data reported as weekly occurrence. Based on these criteria, the screening process resulted in a final sample of 112 client records ($N = 112$).

The rationale for the inclusion/exclusion criteria was based on the following: (1) requirements for the companion thesis (Boutsis, 2016); and (2) eligibility criteria used by the agency for providing services through the ABA-based program being evaluated. For example, as

per requirements of the ABA-based program, only children and adolescents with ASD aged 17-years and younger were included in the sample. Similarly, based on criteria for appropriate statistical analyses used by Boutsis (2016), only files with minimum of three data points were included in the study (Boutsis, 2016).

Measure

The BATT is organized as a checklist including a taxonomy of assessment and behaviour change procedures that are divided into five main sections with several subcategories (see Appendix A). Section 1 covers documentation of anonymized non-identifying data pertaining to the client file such as client ID and target behaviours. Section 2 covers documentation of assessment information including collection of background information, standardized assessment of behaviour/mental health, skills assessments, psychiatric measures, functional behaviour assessment tools and strategies used, and assessment results (e.g., overall hypothesized function). Section 3 covers documentation of behavioural intervention strategies used for preventing challenging behaviours through environmental changes (i.e., antecedent strategies), promoting skill development (i.e., reinforcement), decreasing challenging behaviours through direct intervention (i.e., extinction and punishment), strategies used to train mediators for treatment implementation, and crisis management and safety related recommendations. Section 4 covers data collection methods used to monitor the treatment while Section 5 includes a summary of the case as well as documentation of up to five baseline, treatment, and follow up data points. In the current study, the inclusion criteria involved a minimum of three data points. Hence, if more than three data points were available, only (and up to) the last five data points for each phase of treatment were recorded onto the BATT.

Data Extraction and Abstraction Procedures

For the data extraction phase, a detailed flowchart was developed by the research team (including student investigators for the current thesis and companion thesis) to specify the specific documents to use when retrieving relevant information from the clinical files. The student investigators piloted the flowchart with sample ineligible files and refined the procedures until a reliable method was produced for extracting information from client records (i.e., agreement of 90% or above on sample files). Next, three ineligible test files were used to establish reliability for the two raters with minimum criteria of 90% or above agreement across all three files. Once reliability was established, two additional research assistants were recruited from the participating agency and trained for data collection. The training involved the student investigators providing an overview of the flowchart and the BATT, followed by testing using three test files. Once agreement was established with the research assistants (overall percent agreement of 90% or above across three test files), the file review process was initiated.

Three members of the research team (one student investigator and two research assistants) extracted data for the 112 files using the BATT. This student investigator extracted reliability data for 50 randomly selected files (i.e., 45% of the overall sample). All procedures relating to the sampling and coding of the reliability files were conducted by this student investigator. The random sampling procedure included drawing anonymized client ID numbers for 45% of the files assigned to each rater until a sample of 50 was obtained. In order to ensure continuous monitoring of data quality/reliability, a checkpoint was established at the half-way point with 25% of the reliability sample (i.e., 13 files) being randomly sampled across the three data extractors. Any file with percent agreement below 70% was to be reviewed to identify discrepancies between the extractors however such a situation was never encountered. Next, the

remaining 75% of the reliability sample (i.e., 37 files) was completed. Once the data extraction phase was completed, data abstraction procedures were initiated for the following key variables pertaining to Section 1 (i.e., target behaviours) and Section 3 (i.e., treatment recommendations): (1) target behaviour to increase, (2) target behaviour to decrease, (3) prevention/antecedent strategies, (4) skill building/teaching strategies, and (5) intervention/consequence strategies.

The first step in the data abstraction procedure necessary for retrospective use involved refining the BATT coding scheme. Each item was operationally defined using an introductory Applied Behaviour Analysis textbook (Cooper et al., 2007). Furthermore, other academic sources were also used to assist with breaking down more general concepts (e.g. antecedent interventions) into specific recommendations (e.g., function based interventions, default interventions, etc.; Smith, 2011). The operational definitions were subjected to practice coding using ineligible files. Any discrepancies in coding were discussed and the coding scheme was refined as needed. This process was repeated until sufficient level of agreement was reached on practice files (i.e., 90% or above agreement). Next, reliability testing was completed by the student investigators on four additional ineligible test files with minimum criteria of 90% or above agreement across all four files. Overall percent agreement was calculated following each test file along with discussion of discrepancies amongst the raters prior to introducing the next file.

Once reliability was established across four consecutive files for the student investigators, two additional research assistants were recruited from within Dr. Condillac's lab and trained for coding. The training involved the student investigators providing a detailed overview of the items in the coding scheme, along with the research assistants completing six ineligible practice files each. Once all six practice files were completed and any errors were thoroughly discussed,

the assistants completed reliability testing on four test files with the same minimum criteria of 90% agreement over all four files. Once again, overall percent agreement was calculated after administering each test file and any discrepancies in ratings were discussed prior to introducing the next file. The training for data abstraction was conducted for approximately five hours per rater and was completed over a three day period.

Next, the coding procedure was initiated with three raters coding the 112 extracted files while this student investigator provided a second rating for a sample of 50 files (i.e., the same files that were randomly chosen during extraction). In order to ensure continuous monitoring of data quality/reliability, a checkpoint was established at the half-way point with 30% of the reliability sample (i.e., 15 files) being randomly sampled across the three data coders. Any discrepancies in coding were discussed and addressed with each of the raters. Next, the remaining 70% of the reliability sample (i.e., 35 files) was completed and assessed by this student investigator for reliability. Upon completion of data coding, all data from each of the raters were pooled and entered into IBM SPSS software version 22.0 for analysis.

Results

Data Extraction

As described above, the BATT was completed for a total of 112 client files with a second rater completing the measure for 50 files. The first part of the analysis involved determining the accuracy of data extracted retrospectively, as the BATT was initially designed to be completed prospectively by the clinician doing the work. Percent agreement was calculated for a total of 71 items (overall agreement of 97%; see Table 1), with 21 of the items corresponding to agency required components of the files reviewed. During the training phase for data extraction, overall agreement between the raters on three test files was found to be 97%. For the extracted data,

there were 29 items with 100% agreement that resulted from all non-occurrence data (i.e., zero instances of the item being located in the file by rater 1 or 2; see Table 2). The remaining 41 items had at least one occurrence found by one or both of the raters. Agreement of 90% or above was achieved for 36 of the 41 items, with only one item found to be below 80%. Furthermore, overall agreement for the agency required components was 93%. Prior to coding, the overall agreement for target behaviours and treatment categories was 92% during the extraction phase.

Table 1

Extracted Variables with Occurrence Data

Extracted Item	N	Percent Agreement
(1) Target Behaviours - Overall**	**100**	**88%**
Target Behaviour to Increase	50	84%
Target Behaviour to Decrease	50	92%
(2) Reports Reviewed – Overall	**450**	**100%**
Occupational Therapy	50	98%
(3) School and Medical Status Reviewed - Overall**	**100**	**100%**
School Status	50	100%
Medical Status	50	100%
(4) Medications**	**75**	**93%**
(5) Assessments Used – Overall	**700**	**100%**
Functional Communication Record**	50	98%
Indirect Preference Assessment**	50	100%
Screening Tool for Feeding Problems	50	98%
Social Skills Rating System	50	98%
(6) Functional Behaviour Assessments - Overall	**900**	**96%**
Interview - Behavioural Contextual**	50	98%
Interview - Biopsychosocial**	50	98%
Questionnaire – QABF	50	94%
Questionnaire – FAST	50	100%
Questionnaire – FIDD	50	100%
Other Questionnaires	50	98%
Direct Measure - ABC by Consultant**	50	94%
Direct Measure - ABC by Mediator**	50	74%
Direct Measure - In-Vivo Probe**	50	80%
Direct Measure - ABC Unspecified**	50	88%
Other Direct Measures	50	96%
(7) Overall Hypothesized Function**	**69**	**90%**

(8) Treatment Recommendations - Overall**	**150**	**95%**
Prevention Strategies	50	94%
Skill Building Strategies	50	94%
Intervention Strategies	50	98%
(9) Mediator Training – Overall	**404**	**95%**
BST Model	50	96%
Instruction	50	90%
Modeling – Live	50	96%
Modeling – Video	50	98%
Roleplay	50	92%
Feedback	50	92%
Adherence Measure (with percentage)	54	93%
Other Mediator Training	50	100%
(10) Data Collection and Monitoring	**50**	**94%**
(11) Case Summary – Overall	**200**	**97%**
Goal**	50	98%
Direction of Desired Effect	50	90%
Type of Data Collected	50	100%
Type of Data Graphed	50	100%
(12) Data Points - Overall**	**502**	**91%**
Data Points – Baseline	232	93%
Data Points – Treatment	260	90%
Data Points – Followup	10	100%
(13) Number of Recommendations	**50**	**88%**
(14) Required Agency Items - Overall	**1446**	**93%**
(15) Items to be Coded - Overall	**250**	**92%**
(16) Extraction Training Data – Overall	**213**	**97%**
(17) Extraction Data – Overall	**3950**	**97%**

Note. **Items corresponding to information required by agency to be included in all client files.

Table 2

Extracted Variables with Zero Occurrence Data

Extracted Item	N	Percent Agreement
(1) Reports Reviewed		
Medical	50	100%
Psychiatric	50	100%
Developmental	50	100%
Academic	50	100%
Cognitive	50	100%
Communication	50	100%
Sensory Assessment	50	100%

Other Reports Reviewed	50	100%
(2) Assessments Used		
Adaptive Behaviour	50	100%
Children Sleep Habits Questionnaire	50	100%
Communication (e.g., VB MAPP)	50	100%
Daily Living Skills (e.g., ABLLS)	50	100%
Sexuality	50	100%
Skill Specific (e.g., Toileting)	50	100%
Social Skills	50	100%
Toileting Training Readiness Checklist	50	100%
VB-MAPP (Bathing, Grooming, Dressing)	50	100%
Other Assessments Used	50	100%
(3) Psychiatric Measures		
Dual Diagnosis Screen	50	100%
Severity Scale of Problem Behaviour	50	100%
Specific Diagnostic Screen	50	100%
Quality of Life	50	100%
Other Psychiatric Measures	50	100%
(4) Functional Behaviour Assessments		
Other Interviews	50	100%
Questionnaire – MAS	50	100%
Direct Measure – FAO Form	50	100%
Experimental - Analogue Frequency	50	100%
Experimental - Analogue – Latency	50	100%
Experimental - Structural Analogue	50	100%
Other Experimental Methods	50	100%

Data Abstraction

The second part of the analysis involved calculating the inter-rater reliability for information collected through the BATT that required rater judgement/coding. During the training phase for data abstraction, overall agreement between the raters on four test files was found to be 95%. The remaining analyses were conducted using both percent agreement and Cohen's kappa statistic (κ; Cohen, 1960). Cohen's kappa values can range between 1 and -1 with values close to 1 representing strong agreement. All kappa values in the analysis were interpreted using the conservative criteria specified by Krippendorff (1980) who recommends that only

tentative conclusions be made about data when κ ranges between .67 and .80 while definite conclusions about data should only be reserved for κ above .80.

One of the main concerns with Cohen's kappa is its susceptibility to effects of prevalence and bias (Byrt, Bishop, & Carlin, 1993). The prevalence effect refers to the difference in the proportion of occurrence agreements (i.e., both raters agree on presence of some attribute) and non-occurrence agreements (i.e., both raters agree on absence of the attribute). Similarly, bias refers to the difference in the proportion of occurrence and non-occurrence disagreements between the two raters. A large prevalence index (PI) can inflate chance agreement which in turn can reduce the magnitude of kappa, while a large bias index (BI) can lead to an increase in the magnitude of kappa (Byrt et al., 1993; Sim & Wright, 2005). Although kappa can be corrected for prevalence and bias effects, Byrt et al. (1993) argue against adjusting kappa and simply presenting the PI and BI in addition to the non-adjusted kappa. Hence, prevalence and bias indices for each of the variables are reported along with the kappa value.

In addition, 95% confidence intervals (95% CI) corresponding to each kappa score were calculated using the standard error (SE) estimate and corresponding z-score with the formula: $\kappa \pm 1.96*SE$. The purpose of calculating confidence intervals was to create a measure of accuracy for kappa by providing a range of values that contain the true population value for any given kappa (Sim & Reid, 1999). Furthermore, the lower limit of the confidence interval can be used as a tool for evaluating clinical significance by comparing it to the minimum acceptable magnitude of the test statistic (Sim & Reid, 1999). In this study, any variable with the lower limit of the confidence interval falling below .67 was treated as unreliable based on the threshold specified by Krippendorff (1980).

Furthermore, any items with low frequency of observations ($n < 10$) were combined with other items whenever conceptually permissible to create a minimum sample size of 10 ratings in order to include the item in the reliability analysis. This minimum sample size calculation was based on requirements for a 1-tailed test (null of $\kappa = 0$) with 80% power to detect a statistically significant kappa of .80 ($p \leq .05$) for two raters on a dichotomous variable (Sim & Wright, 2005). A 1-tailed test was used instead of a two-tailed test as the null hypothesis for testing kappa included a value of zero and the only meaningful interpretation of negative kappa values is that the level of observed agreement is based on chance alone (Brennan & Silman, 1992).

Lastly, there were a number of items without any observed data. Given that no conclusions can be drawn about such items, the decision was made to exclude all such items from the reliability analysis (see Table 3 for list of items without any sampled data).

Table 3

Items Excluded from Analysis

Item Abstracted	N
Target Behaviour to Decrease	
Dangerous Behaviours	
Pica	0
Vomiting/Regurgitation	0
Inappropriate Communication	
Self-Talk	0
Aberrant Speech	0
Physical Stereotypy	0
Non-Specified Behaviour	0
Off Task Behaviour	0
Inappropriate Sexual	
Stripping	0
Inappropriate Touch Others	0
Public Masturbation	0
Miscellaneous	
Fecal Smearing	0
Urine/Bowel Accident	0
Target Behaviours to Increase	
Homework	0

Self-Regulation	0
Communication-Other	0
Treatment Recommendations	
Skill Building	
Skill Building-Generalization or Maintenance	0
Skill Building-ADL	0
Skill Building+Extinction	0
Intervention	
Intervention+Prevention	0
Intervention-Negative Punishment	0
Intervention-Positive Punishment	0
Intervention-Response Blocking	0
Intervention-Combined with Punishment	0

Missing data. Any instances where one rater failed to provide a rating were counted as a disagreement. Missing data can have unintended effects on the magnitude of kappa depending on the proportion of data missing and the pattern of missingness (Adejumo, 2005). Hence, missing data were assigned a separate code and included in the reliability analysis to reduce any unintended effects. Proportion of missing data per variable are described in Table 4.

Table 4

Proportion of Missing Data for Coded Variables

Item Abstracted	N	Proportion Missing
Target Behaviour to Decrease	121	1.70%
Target Behaviour to Increase	47	0.00%
Types of Recommendations	467	2.14%
Proactive versus Reactive	467	2.14%
Client versus Mediator	467	2.14%

Target behaviours to decrease. The BATT coding scheme for target behaviour to decrease involved 7 categories of items split into a total of 30 codes. The overall kappa prior to combining codes was .95 (95% CI [.91, .99], $p < .001$) while overall percent agreement was 95.87%. Any codes within each category with less than 10 ratings were combined to increase sample size of the items being analyzed. This process resulted in a final set of 7 codes (see Table

5). Furthermore, all 7 codes achieved excellent reliability with $\kappa > .90$ and percent agreement above 95% for each of the items.

Table 5

Reliability Data for Combined Target Behaviour to Decrease

Item Abstracted	N	Percent Agreement	κ (95% CI)*	PI	BI
Dangerous behaviours	22	99.17%	0.97 (0.92-1.00)	0.63	0.01
Aggression	18				
Self-Injury	3				
Destruction	1				
Inappropriate Communication	20	97.52%	0.91 (0.81-1.00)	0.68	0.01
Aggressive Threats	2				
Swearing	3				
Name Calling	9				
Repeated Questions	6				
Refusal/Protest	14	99.17%	0.96 (0.89-1.00)	0.76	0.01
Disruptive - Crying	12	100%	1.00 (1.00-1.00)	0.80	0.00
Disruptive - Screaming	16	99.17%	0.96 (0.89-1.00)	0.74	0.01
Disruptive – Other	13	100%	1.00 (1.00-1.00)	0.79	0.00
Whining	5				
Tantrums	4				
Throwing	4				
Miscellaneous Behaviours	24	98.35%	0.95 (0.87-1.00)	0.62	0.02
Taking or Grabbing Items	3				
Flopping	2				
Running	7				
Verbal Stereotypy	1				
Other	11				

Note. κ = Cohen's kappa; CI = Confidence Interval; PI = Prevalence Index; BI = Bias Index.
* $p < .001$

Target behaviours to increase. Coding scheme for behaviours to increase included 13 categories involving a total of 17 different codes. The overall kappa prior to combining codes was .90 (95% CI [.78-1.00], $p < .001$) while overall percent agreement was 93.62% (see Table 6). Similar to the steps taken with behaviours to decrease, items with low frequency of ratings

were grouped together based on conceptual similarity. This process resulted in the following 3 codes: cooperation (n = 11), tolerance/waiting (n = 10), and other adaptive behaviours (n = 27). Agreement on combined codes was found to be $\kappa > .87$ indicating excellent reliability on these items.

Table 6

Reliability Data for Combined Target Behaviour to Increase

Item Abstracted	N	Percent Agreement	κ (95% CI)*	PI	BI
Cooperation	11	100%	1.00 (1.00-1.00)	0.53	0.00
Tolerance and Waiting	10	97.87%	0.93 (0.81-1.00)	0.60	0.02
Tolerance	5				
Waiting	5				
Other Adaptive Behaviours	26	93.62%	0.87 (0.73-1.00)	0.09	0.02
Transitions	2				
Routine/Schedule	3				
Sharing	1				
Play Skills	4				
Activities of Daily Living (ADL)	3				
Safety Skills	2				
Self-Management	1				
Communication Mands	3				
Communication Tacts	1				
Communication-Intraverbals	2				
Other	4				

Note. κ = Cohen's kappa; CI = Confidence Interval; PI = Prevalence Index; BI = Bias Index.
* $p < .001$

Treatment recommendations. The initial BATT coding scheme for treatment recommendations included 3 main categories: prevention, skill building, and intervention. The prevention category included 12 codes, skill building comprised of 12 codes, and intervention involved 13 codes for a total of 37 codes representing different treatment recommendations. Initial attempts to determine reliability involved calculating kappa for the overall variable

including all 37 codes. This analysis revealed inter-rater reliability below the acceptable threshold (κ = .62, 95% CI [.57, .67], p < .001). Next, reliability was calculated per treatment section including all corresponding codes. This analysis also revealed unacceptable level of reliability for prevention (κ = .51 [.43, .59], p < .001), skill building (κ = .56 [.47, .64], p < .001), and intervention (κ = .66 [.55, .78], p < .001) sections. Given these results, the next analysis involved determining reliability for the treatment sections at the categorical level (i.e., all within-category codes collapsed), which revealed excellent overall reliability for prevention (κ = .84 [.79, .89], p < .001), skill building (κ = .81 [.76, .87], p < .001), and intervention (κ = .90 [.84, .95], p < .001) sections. Based on the results achieved, the next step involved analysing specific recommendations within each of the treatment sections. A number of items had low sample sizes (i.e., n < 10) and so a decision was made based on criteria identified above to combine conceptually similar items within each of the treatment sections to increase sample size. This process resulted in a total of 12 codes: three codes for the prevention category (see Table 7), five codes for the skill building category (see Table 8), and four codes for the intervention category (see Table 9) categories. However, at the level of specific recommendations, only five of the 12 items achieved κ > .67 and only one item (i.e., Intervention-Combined) achieved κ > .80.

Table 7

Treatment Recommendations - Prevention

Item Abstracted	N	Percent Agreement	κ (95% CI)*	PI	BI
Prevention Category	**195**	**89.50%**	**0.84 (0.79-0.89)**	**0.16**	**0.01**
Prevention-Alter SD/MO	162	87.79%	0.73 (0.66-0.80)	0.32	0.01
Prevention-Alter/Remove Sd for identified Problem Bx	13				

Prevention-Alter/Add/ Sd to evoke an alternative Bx	76				
Prevention-AO for Problem Bx	21				
Prevention-NCR to Reduce Problem Bx	2				
Prevention-DRO to Reduce Problem Bx	2				
Prevention-Set Expectations	48				
Prevention-Default Strategies	13	97.79%	0.43 (0.19-0.67)	0.94	0.00
Prevention-Other	20	96.57%	0.65 (0.49-0.81)	0.90	0.02
Prevention-Adherence	1				
Prevention-Alter Transition	2				
Prevention-Biomedical	4				
Prevention-Safety	5				
Prevention-Combined	8				

Note. κ = Cohen's kappa; CI = Confidence Interval; PI = Prevalence Index; BI = Bias Index.
* $p < .001$

Table 8

Treatment Recommendations – Skill Building

Item Abstracted	N	Percent Agreement	κ (95% CI)*	PI	BI
Skill Building Category	**172**	**85.88%**	**0.81 (0.76-0.87)**	**0.26**	**0.00**
Skill Building-Reinforcement Procedure non-specific	21	97.22%	0.67 (0.50-0.84)	0.91	0.00
Skill Building-Communication (Function Related)	33	96.57%	0.72 (0.60-0.85)	0.87	0.01
Skill Building-Other Replacement Skill (Function Related)	68	90.36%	0.66 (0.57-0.75)	0.66	0.04
Skill Building-Non-Function Related	35	94.65%	0.58 (0.43-0.72)	0.87	0.01
Skill Building-Communication (Non-Function Related)	6				
Skill Building-Other Skill (Non-Function Related)	30				
Skill Building - Other	15	96.57%	0.26 (0.01-0.51)	0.95	0.02
Skill Building-Adherence	1				

Skill Building-DRL / DRH	6
Skill Building-Combined	1
Combine-Prevention+Skill Building	7

Note. κ = Cohen's kappa; CI = Confidence Interval; PI = Prevalence Index; BI = Bias Index.
* $p < .001$

Table 9

Treatment Recommendations – Intervention & Proactive vs Reactive

Item Abstracted	N	Percent Agreement	κ (95% CI)*	PI	BI
Intervention Category	**90**	**90.0%**	**0.90 (0.84-0.95)**	**0.62**	**0.01**
Intervention-Redirection	29	96.57%	0.72 (0.58-0.85)	0.87	0.00
Intervention-Minimize Response (non-function)	15	98.50%	0.73 (0.54-0.92)	0.94	0.01
Intervention-Combined no Punishment	34	97.88%	0.83 (0.73-0.93)	0.86	0.01
Intervention – Other	12	98.29%	0.68 (0.48-0.89)	0.94	0.00
Intervention-Adherence	1				
Intervention-Safety	4				
Intervention-Attenuate Reinforcement (function related)	4				
Intervention-Extinction (function related)	2				
Intervention-Follow Through	1				
Client versus Mediator	**467**	**96%**	**N/A**		
Proactive versus Reactive	**467**	**93.36%**	**0.80 (0.74-0.87)**	**0.60**	**0.01**

Note. κ = Cohen's kappa; CI = Confidence Interval; PI = Prevalence Index; BI = Bias Index.
* $p < .001$

Given the low level of agreement found for specific recommendations, a case-study was conducted by isolating one of the problematic items in order to identify a potential source of disagreement. As displayed in Table 7, the item "Prevention-Alter SD/MO" was formulated by combining data for six items based on conceptual similarity and sample sizes. Table 10 describes the cross-tabulation of the frequency of ratings provided by Rater 1 and Rater 2 across each of the items being analysed. Based on data in Table 10, there was some overlap discovered between ratings for Rater 1 and 2 for the specific recommendations: For example, 50% (n = 9) of ratings

for Rater 1 on Item 3 were found to overlap with Item 2 for Rater 2. Similarly, 25% (*n* = 18) of ratings for Rater 2 on Item 2 were found to overlap with Items 1, 3, and 6 for Rater 1. Therefore, a potential source of error in the reliability data (which resulted in low agreement) was the overlap found between the definitions of individual treatment recommendations.

Table 10

Frequency of Ratings per Item for One Combined Variable for Rater 1 and 2

	RATER 2						
RATER 1	**1**	**2**	**3**	**4**	**5**	**6**	**Total**
1 Prevention-Alter/Remove SD	5	1	1	0	0	0	**7**
2 Prevention-Alter/Add SD	0	57	2	0	0	2	**61**
3 Prevention-AO for Problem Bx	0	9	9	0	0	0	**18**
4 Prevention-NCR	0	0	0	1	0	0	**1**
5 Prevention-DRO	0	0	0	0	1	0	**1**
6 Prevention-Set Expectations	0	8	0	0	0	35	**43**
Total	**5**	**75**	**12**	**1**	**1**	**37**	**131**

Note. Grey highlights indicate exact agreement between Rater 1 and Rater 2 for individual items.

In order to further investigate the discrepancy identified, reliability data from the data abstraction testing phase were re-evaluated for each of the raters for the specific treatment recommendations. As displayed in Table 11, overall agreement across the raters for the specific recommendations was found to be 85%, which was below the 90% criteria established during testing. Furthermore, analysis of the four files used for testing rater agreement revealed that the raters were exposed to only seven (out of 12) prevention codes, four (out of 14) skill building codes, and two (out of 11) intervention codes during training.

Table 11

Percent Agreement Across Raters During the Testing Phase

Rater	**Percent Agreement - Specific Recommendations**
Rater 1A - Student Investigator	85%
Rater 1B - Research Assistant 1	85%
Rater 1C - Research Assistant 2	76%

Rater 2 – This Student Investigator	93%
Overall	**85%**

Client versus mediator. Another coded variable included rating whether the recommendation was written for the mediator (e.g., care provider) or the client to implement (i.e., individual receiving treatment). For this analysis, percent agreement was used as the primary measure. Kappa could not be calculated as Rater 2 did not rate any of the recommendations as being written for the client to implement, while only one recommendation was rated as being client implemented by Rater 1, indicating that recommendations were primarily rated by both raters as mediator implemented. Overall agreement for the client versus mediator categorization was found to be 96%.

Proactive versus reactive. The last variable of analysis involved rating each of the recommendations as either proactive (occurring prior to the target behaviour) or reactive (immediately following the target behaviour). Based on these definitions, recommendations categorized as prevention and skill building were hypothesized to fall in the proactive category while reactive category was hypothesized to correspond with recommendations categorized as intervention. For this analysis, a sufficiently large sample size (N = 467) resulted in excellent reliability with kappa of .80 (95% CI [0.74, 0.87], $p < .001$) and percent agreement of 93.66% (see Table 8).

Post Hoc Internal Reliability Analysis

As reliability was established for the prevention, skill building, and intervention sections at a categorical level, but not at the individual treatment level, further post-hoc analysis involved assessing the internal reliability of the treatment section codes of the BATT (i.e., Prevention, Skill Building, and Intervention). The treatment categories for Rater 1 were compared with the

rating provided by Rater 2 regarding the proactive versus reactive nature of the recommendation for the 50 cases assessed for reliability. The analysis was performed after removing cases with missing ratings from Rater 1 or Rater 2, as missing codes were already captured in the inter-rater reliability analyses. The use of ratings provided by Rater 2 was included in order to address the issue of non-independence that would exist if both ratings corresponded to Rater 1. Overall, the analysis revealed excellent inter-rater reliability between the ratings provided by Rater 1 for the treatment sections and the ratings provided by Rater 2 on the proactive versus reactive nature of the recommendations (κ = .87, 95% CI [.82, .91], p < .001), suggesting internal reliability of the treatment section codes.

Discussion

The purpose of the current investigation was to test the inter-rater reliability of the BATT through retrospective review of 112 client case records. The files were obtained from an agency providing ABA-based behavioural services to individuals with a diagnosis of ASD. The study had three main purposes: (1) test the retrospective utility of a tool designed for prospective use; (2) pilot a tool that can assist in the comprehensive documentation of assessment and treatment strategies used by clinicians; and (3) to create a standardized data gathering system that may be utilized for program evaluation within and across ABA-based treatment programs. Furthermore, there were two main research questions that were tested in this study using separate analyses.

The first set of analyses involved determining the accuracy of the 71 extracted variables by calculating percent agreement across three categories of items: (1) items corresponding to information required by the agency to be included in all files, (2) other relevant information that is consistent with best practices, and (3) items that were later coded using the BATT coding

scheme. The results indicated excellent agreement (i.e., 90% or above) for all three categories, with only five of the 71 extracted variables achieving agreement below 90%.

The second set of analyses involved testing the inter-rater reliability of five different coded nominal variables with various subcategories. The results revealed excellent inter-rater agreement (i.e., $\kappa > .80$) on the following: (1) target behaviours to decrease, (2) target behaviours to increase, (3) treatment recommendations (at the categorical level), (4) the use of proactive versus reactive strategies; and (5) client versus mediator implemented recommendations. The exception to these results was the coding of specific strategies within the prevention, skill building, and intervention sections.

Post hoc analyses were conducted to determine if any explanations for the low reliability on specific treatment codes existed. First, the data from coder training were examined, which revealed that lower reliability scores on the specific strategies were masked by excellent reliability on other items as overall percentage agreement of the treatment related codes was calculated to assess reliability (instead of reliability per item).An examination of the files that were tested during the training phase revealed that the selected test files sampled only 13 out of 37 codes which was fewer codes than anticipated, the implications of which are discussed in the limitations section below.

Lastly, given the acceptable inter-rater reliability achieved for the larger treatment categories, further post hoc analysis was performed to assess the internal reliability of the treatment categories. A comparison was made between ratings provided by Rater 1 on the overall treatment sections with ratings provided by Rater 2 on the proactive versus reactive nature of the recommendations. Overall, the results revealed excellent inter-rater reliability between the

ratings for the two variables ($\kappa > .80$) suggesting that the larger treatment categories could be used in further analyses for the companion thesis (Boutsis, 2016).

Research Question 1: Can the BATT be used retrospectively with a high degree of reliability to extract necessary information for program evaluation?

According to Baker et al. (2014), the competency of the clinician and the integrity with which mediators adopt recommendations is an essential component of evaluating mediator-implemented interventions. Furthermore, the measurement and reporting of treatment integrity data are generally lacking in outcome research (McIntyre et al., 2007; Wheeler et al., 2006). Although the BATT was originally designed for prospective use for program evaluation of ABA treatments, the purpose of this investigation was to evaluate its utility in a retrospective study design. The BATT includes assessment, treatment, and monitoring strategies and their outcomes. The information extracted from the files contained components required by the agency for all cases as well as additional relevant information that is consistent with best practices. Based on the results described above, an excellent level of agreement was achieved during the extraction phase. The BATT items corresponding to required components of the file and other relevant information were found to be highly reliable. This finding demonstrates the sensitivity of the BATT for retrospectively detecting between-clinician variability in the types of assessment and treatment tools utilized by clinicians working at the agency. Hence, the BATT can serve a useful purpose for ABA-based agencies to retrospectively evaluate the training needs of their clinicians by capturing information on the types of tools they utilize in their practice. This type of analysis can be used as the first step in the larger goal of wide-scale program evaluation within and across ABA-based service providers.

Research Question 2: Can the BATT coding scheme be used with high degree of reliability to code key variables for program evaluation?

Given the acceptable level of reliability achieved in the extraction of data from agency files, the next phase involved using the BATT coding scheme to code information for the purpose of program evaluation. The results indicated excellent inter-rater agreement for all coded variables including target behaviours, the overall treatment sections, identification of client versus mediator-implemented recommendations, and the identification of proactive versus reactive strategies.

Although reliability was not demonstrated at the level of specific recommendations, the results of this study have important implications for program evaluation and outcome research in the field of ABA. The purpose of the BATT is to help standardize documentation practices used by ABA-based service providers. The coding scheme included on the BATT provided an effective method for categorizing multi-component target behaviours (e.g., multiple topographies included in one operational definition) and complex multi-step treatment recommendations.

Furthermore, the identification of the nature of recommendations as proactive versus reactive can allow ABA-based agencies to evaluate the restrictiveness of the treatments employed. In their review of behavioural intervention research, Brosnan and Healy (2011) identified the efficacy of antecedent-based interventions as a least-restrictive alternative to punishment based or other behaviour reduction procedures (e.g., extinction). Although the authors noted that intrusive measures were used in a number of the studies reviewed, they advocated for the use of more preventative strategies prior to the use of more intrusive measures (Brosnan & Healy, 2011). The results of the current study suggest that the BATT can be used

retrospectively to identify the nature of interventions used by clinicians and this information can be used by ABA-based agencies to evaluate the restrictiveness of their treatments.

Lastly, the client versus mediator categorization also produced reliable results. This variable was included for evaluating whether recommendations were written for the mediator or the client to implement. From the data analysed in this study, nearly all ratings were based on the mediator implementing the recommendations (hence why Kappa could not be calculated). This was not surprising given that the data evaluated in this study pertained to an ABA-based program using the mediator model. The inclusion of this variable provided further confirmation that the strategies used by clinicians were mediator focused and our evaluation was centered on a mediator-based approach for treating challenging behaviours. For ABA-based agencies employing the mediator model, this information can be helpful for evaluating the prevalence of recommendations aimed at teaching skills to mediators (i.e., feature of the mediator model) in comparison to skills taught directly to the client.

Strengths

The primary purpose for developing the BATT was to establish a standardized method of documenting assessment and treatment components of behavioural interventions prospectively. This study established the reliability of a two-step process for data collection (i.e., extraction and abstraction). During the extraction phase, research assistants completed the BATT by summarizing factual information directly onto the form. No personal identifying information was copied onto the BATT. This minimized the risk of accidental breach of private information during the abstraction phase when multiple coders needed to access the files. The results of this study provided preliminary evidence of the internal reliability of the BATT coding scheme used for categorizing recommendations across the overall treatment sections.

There are a number of advantages of collecting information in a standardized format. Firstly, there has been limited focus in the past on identifying the active ingredients likely responsible for the changes in behaviour especially for more complex behaviour change programs (Abraham & Michie, 2008; Michie et al., 2013). Although there is sufficient evidence available for effectiveness of specific ABA procedures, many intervention programs generally include a mix of prevention, skill building, and intervention strategies (Feldman et al., 2002). In the companion study, there were an average of 9.3 recommendations documented per file across the 112 cases reviewed, which permitted analysis of the effectiveness of the combined treatment package (Boutsis, 2016).

However, there is a paucity of research examining the interaction between level of clinician competency (e.g., education, experience, etc.) with client and mediator characteristics (e.g., severity of behaviour, type of training received, etc.). The BATT is designed to summarize information on a number of aspects of the assessment and treatment process including: (a) types of assessments reviewed by the clinician, (b) use of psychotropic medications, (c) behavioural assessment methods used including indirect measures, descriptive assessments, and functional analyses, (d) types of prevention, skill building, and intervention strategies utilized, and (e) information regarding mediator training procedures. Hence, the BATT can allow for a comprehensive evaluation of ABA-based treatment programs within and across service providers.

Implications for Practice

Currently, there are no published systems that could be used to sufficiently capture information for the purpose of wide-scale program evaluation of ABA services for children with ASD. Although each agency has documentation standards and information is kept on file for

each of the clients serviced, the variability in documentation within and across service agencies presents a significant barrier to this type of analysis. This is precisely where the BATT may be able to serve a useful function. By offering a standardized method of documenting and monitoring behavioural interventions, the BATT can successfully extend the assessment and treatment literature (as well as clinical practice) in a number of ways. Firstly, the BATT can be used by behaviour change agencies to develop standards for treatment integrity pertaining to services delivered directly or indirectly by therapists. For example, agencies can require behaviour therapists to document specific assessment methods used and evaluate whether therapists meet the requirements of evidence-based practices outlined in the behaviour analysis literature (e.g., use of experimental functional analyses). Secondly, the BATT can be used as a tool for formally monitoring interventions on an ongoing basis. This can allow clinical supervisors to track and evaluate the types of treatment strategies utilized by therapists and to identify strengths and gaps in the clinical training of the therapists.

Other benefits of utilizing the BATT include the assessment of treatment outcomes for individual clients as well as large-scale program evaluation across multiple therapists and sites. The BATT can serve as a guide for quantifying, categorizing, and coding data contained in closed files. In addition, ABA-based treatment programs can use this tool to demonstrate the efficacy of their services to funding agencies in order to secure more funds and broaden the scope of their services.

Furthermore, researchers may also be able to use these data to analyse the types of behavioural strategies used in clinical practice and the interactive nature of different behaviour change techniques when used in combination. This may enable practitioners to maximize the efficacy of multi-component behavioural interventions by introducing techniques that work best

independently and in combination. Lastly, data on the types of strategies used may also generate information about the types of procedures that are seldom employed in clinical practice (e.g., positive punishment). This may be important information for the field of behaviour analysis, which is based on the premise of implementing the least-intrusive methods of intervention (Brosnan & Healy, 2011).

Limitations and Future Research Considerations

There are several limitations that must be taken into consideration when interpreting the results of this study and for future studies utilizing the BATT. Firstly, although there was sufficient sample size for evaluating the abstraction of target behaviours and treatment recommendation sections, there was insufficient sample size for evaluating item specific reliability within each of the treatment sections. One solution would have been to sample more data based on frequencies of the items in order to obtain greater sample size and a more balanced distribution of ratings. Hence, future studies using the BATT should consider basing the sample size on frequency of items of interest.

Secondly, the level of agreement found for specific recommendations was below the specified threshold for nearly all items. In order to investigate the source of disparity observed, post-hoc analyses were conducted. The first analysis revealed that there was significant overlap between items within the treatment sections, indicating that many of the items were not mutually exclusive. Hence, future studies employing the BATT should consider using the combined codes included in the present study and further refinement should be performed prior to beginning data abstraction.

The second analysis helped identify issues with the training system developed for the raters on data abstraction. The criteria used for assessing rater reliability during the testing phase

was based on overall agreement of 90% and above across all variables. This implicates that reliability on the specific recommendations was not calculated separately and post-hoc analysis revealed the overall agreement amongst the raters was only 85% during training. Furthermore, analysis of files used for testing revealed that insufficient exposure was provided to the raters on coding specific items as only a small number of codes per section were officially tested with the test files. Therefore, future studies should assess agreement on specific key variables during testing and consideration should be given to using multiple exemplars for each coded item during the testing phase.

In addition to the challenges identified with the data abstraction training system, another limitation includes the generalizability of findings of this study. The data used in this study was based on a single ABA-based treatment program with specific eligibility criteria. Therefore, generalizability of findings are limited to data analysed for individuals with ASD receiving treatment through a short-term (i.e., 12-week) service model with focus on decreasing challenging behaviours (rather than focusing on other characteristic traits of ASD; e.g., social skills training). Data collection in a more diverse ABA program serving children and adults, as well as individuals with various diagnoses (e.g., Fetal Alcohol Spectrum Disorder; Down Syndrome; dual diagnoses, etc.) would have lead to a more representative sample of individuals with developmental disabilities. Therefore, future studies should consider testing the reliability of the BATT across various ABA-based programs (with varied eligibility criteria).

Furthermore, establishing validity of the data utilized through the BATT is an important goal and future studies should consider the steps necessary to achieve valid results. As mentioned previously, the BATT is a tool designed for documenting behavioural interventions. The BATT can be used for gathering data on a number of aspects of the assessment and

intervention process. Data collected through the BATT may be used for various aspects of program evaluation including the assessment of therapist competence and treatment integrity. Hence, validating data generated through the BATT requires the establishment of: (1) reliability of the items on the BATT; (b) correlation between the results generated from the BATT with another independent assessment of the construct of interest (i.e., convergent validity); and (3) a lack of correlation with other unrelated constructs (i.e., divergent validity; Ryan, Leong, & Oswald, 2012).

Lastly, the retrospective design used in this study had a number of inherent limitations that may have also impacted the results. As mentioned previously, the BATT was designed for prospective use and the purpose of this study was to test its utility for retrospective program evaluation. One of the main challenges involved with retrospectively reviewing files is the lack of appropriate contextual information available to make accurate judgements. For example, some of the terminology encountered in the documentation involved non-behavioural or ambiguous language (e.g., “avoid power struggles”) making some items more subjective than others and more likely to fall into more than one category of codes. Furthermore, some files had missing information or inconsistent patterns of documentation. This finding may be related to the fact that the inclusionary criteria spanned a 2.5 year period, during which time several changes may have been introduced to the agency’s documentation practices.

Overall, a number of the challenges encountered in this study were artifacts of the study design (i.e., retrospective review) and future studies should consider employing the BATT prospectively. The goal of the current evaluation was to establish inter-rater reliability of the tool for retrospective review. However, prospective use of the BATT may yield better results. Prospective evaluation would involve clinicians using the BATT as a guide to document their

interventions. Furthermore, because the clinician working on a particular file would be involved in filling out the BATT themselves, this would allow for the preservation of contextual information and avoid the ambiguities encountered in the current study. Although the current study produced reliable results in a number of areas, factors related to the retrospective review process adversely impacted the assessment of inter-rater reliability. In this sense, this study represents the lower limit of reliability for the BATT as prospective evaluation is likely to yield better results.

Conclusion

The current study serves an important first step in the retrospective use of the BATT as a program evaluation tool. Acceptable levels of reliability were demonstrated on many components of the BATT with the exception of the treatment recommendations at an item-specific level. Despite the inherent limitations of the study design, the results of this study are encouraging for retrospective evaluation of ABA-based behavioural interventions. Future considerations include addressing sample size issues, modifying training methods, and use of a prospective design for testing the reliability of the BATT for the purpose of program evaluation.

References

Abraham, C., & Michie, S. (2008). A taxonomy of behavior change techniques used in interventions. *Health Psychology, 27*(3), 379-387.

Adejumo, A. O. (2005). Effect of missing values on the Cohen's kappa statistic for raters agreement measurement. *International Journal of Pure and Applied Mathematics, 22*(1), 13-31.

Aman, M. G., Singh, N. N., Steward, A. W., & Field, C. J. (1985). The aberrant behavior checklist: a behavior rating scale for the assessment of treatment effects. *American Journal of Mental Deficiency, 89*(5), 485-91.

American Psychiatric Association. (2013). *Diagnostic and statistical manual of mental disorders* (5th ed.). Washington, DC: Author.

Baer, D. M., Wolf, M. M., & Risley, T. R. (1987). Some still-current dimensions of applied behavior analysis. *Journal of Applied Behavior Analysis, 20*(4), 313–327.

Baker, B. L., McIntyre, L. L., Blacher, J., Crnic, K., Edelbrock, C., & Low, C. (2003). Pre-school children with and without developmental delay: behaviour problems and parenting stress over time. *Journal of Intellectual Disability Research, 47*(5), 217-230.

Baker, C. N., Leff, S. S., Bevans, K., & Power, T. J. (2014). Treatment integrity in urban, community-based prevention programs. In L. M. Hagermoser Sanetti, T. R. Kratochwill, L. M. Hagermoser Sanetti, T. R. Kratochwill (Eds.), *Treatment integrity: A foundation for evidence-based practice in applied psychology* (pp. 279-301). Washington DC: American Psychological Association.

Beavers, G. A., Iwata, B. A., & Lerman, D. C. (2013). Thirty years of research on the functional analysis of problem behavior. *Journal of Applied Behavior Analysis, 46*(1), 1-21.

Biggs, B. K., Vernberg, E. M., Twemlow, S. W., Fonagy, P., & Dill, E. J. (2008). Teacher adherence and its relation to teacher attitudes and student outcomes in an elementary school-based violence prevention program. *School Psychology Review, 37*(4), 533-549.

Boutsis, E. (2016). *A retrospective evaluation of an applied behaviour analysis community based service for children with autism spectrum disorders.* Unpublished master's thesis. Brock University, St. Catharines.

Brennan, P., & Silman, A. (1992). Statistical methods for assessing observer variability in clinical measures. *British Medical Journal, 304*(6840), 1491-1494.

Brosnan, J., & Healy, O. (2011). A review of behavioral interventions for treatment of aggression in individuals with developmental disabilities. *Research in Developmental Disabilities, 32,* 437-446.

Brown, R. I., MacAdam-Crisp, J., Wang, M., & Iarocci, G. (2006). Family quality of life when there is a child with a developmental disability. *Journal of Policy and Practice in Intellectual Disabilities, 3*(4), 238-245.

Byrt, T., Bishop, J., & Carlin, J. B. (1993). Bias, prevalence, and kappa. *Journal of Clinical Epidemiology, 46*(5), 423-429.

Carr, E. G., & Durand, V. M. (1985). Reducing behavior problems through functional communication training. *Journal of Applied Behavior Analysis, 18,* 111-126.

Carr, E. G., Horner, R. H., Turnbull, A. P., Marquis, J. G., McLaughlin, D. M., McAtee, M. L., … & Braddock, D. (1999). *Positive behavior support as an approach for dealing with problem behavior in people with developmental disabilities: A research synthesis.* Washington, DC: AAMR.

Cohen, J. A. (1960). Coefficient of agreement for nominal scales. *Educational and Psychological Measurement, 20*(1), 37–46.

Condillac, R. A. (2009a). *Evaluating processes and outcomes for behaviour therapy services: Phase 1 - Developing the Measures.* Research Proposal Submitted to Brock University Research Ethics Board, St. Catharines, ON: Brock University.

Condillac, R.A. (2009b). *Impact on Quality of Life (IQOL)*. Unpublished Measure. St. Catharines, ON: Brock University.

Condillac, R.A. (2009c). *The Behavioural Assessment and Treatment Taxonomy (BATT).* Unpublished Measure. St. Catharines, ON: Brock University.

Condillac, R.A. (2009d). *The Target Behaviour Severity Scale (TBSS).* Unpublished Measure. St. Catharines, ON: Brock University.

Cooper, J. O., Heron, T. E., & Heward, W. L. (2007). *Applied Behavior Analysis – Second Edition.* Upper Saddle River, New Jersey: Pearson.

Courtemanche, A. B., Schroeder, S. R., & Sheldon, J. B. (2011). Designs and analyses of psychotropic and behavioral interventions for the treatment of problem behavior among people with intellectual and developmental disabilities. *American Association on Intellectual and Developmental Disabilities, 116*(4), 315-328.

DiGennaro, F. D., Martens, B. K., & Kleinmann, A. E. (2007). A comparison of performance feedback procedures on teachers' treatment implementation integrity and students' inappropriate behavior in special education classrooms. *Journal of Applied Behavior Analysis, 40,* 447–461.

DiGennaro Reed, F. D., Reed, D. D., Baez, C. N., & Maguire, H. (2011). A parametric analysis of errors of commission during discrete-trial training. *Journal of Applied Behavior Analysis, 44,* 611–615.

Durand, V. M., & Crimmins, D. G. (1988). Identifying the variables maintaining self-injurious behavior. *Journal of Autism and Developmental Disorders, 18,* 99–117.

Eikeseth, S., Smith, T., Jahr, E., & Eldevik, S. (2002). Intensive behavioral treatment at school for 4-to 7-year-old children with autism a 1-year comparison controlled study. *Behavior Modification*, *26*(1), 49-68.

Feldman, M. A., Condillac, R. A., Tough, S., Hunt, S., & Griffiths, D. (2002). Effectiveness of community positive behavioral intervention for persons with developmental disabilities and severe behavior disorders. *Behavior Therapy, 33,* 377-398.

Fryling, M. J., Wallace, M. D., & Yassine, J. N., (2012). Impact of treatment integrity on intervention effectiveness. *Journal of Applied Behavior Analysis, 45,* 449-453.

Hanley, G. P., Iwata, B. A., & McCord, B. E. (2003). Functional analysis of problem behaviour: A review. *Journal of Applied Behavior Analysis, 36,* 147-185.

Harvey, S. T., Boer, D., Meyer, L. H., & Evans, I. M. (2009). Updating a meta-analysis of intervention research with challenging behaviour: Treatment validity and standards of practice. *Journal of Intellectual & Developmental Disability, 34,* 67–80.

Hayes, S. C., Nelson, R. O., & Jarrett, R. B. (1987). The treatment utility of assessment: a functional approach to evaluating assessment quality. *American Psychologist*, *42*(11), 963-974.

Heyvaert, M., Saenen, L., Campbell, J. M., Maes, B., & Onghena, P. (2014). Efficacy of behavioral interventions for reducing problem behavior in persons with autism: An updated quantitative synthesis of single-subject research. *Research in Developmental Disabilities, 35,* 2463-2476.

Ingersoll, B., & Dvortcsak, A. (2006). Including parent training in the early childhood special education curriculum for children with autism spectrum disorders. *Journal of Positive Behavior Interventions, 8*(2), 79–87.

Iwata, B. A., Dorsey, M. F., Slifer, K. J., Bauman, K. E., & Richman, G. S. (1994). Toward a functional analysis of self-injury. *Journal of Applied Behavior Analysis, 27,* 197–209. (Reprinted from *Analysis and Intervention in Developmental Disabilities, 2,* 1982, 3–20).

Jocelyn, L. J., Casiro, O. G., Beattie, D., Bow, J., & Kneisz, J. (1998). Treatment of children with autism: a randomized controlled trial to evaluate a caregiver-based intervention program in community day-care centers. *Journal of Developmental and Behavioral Pediatrics, 19*(5), 326-334.

Kaale, A., Smith, L., & Sponheim, E. (2012). A randomized controlled trial of preschool-based joint attention intervention for children with autism. *Journal of Child Psychology and Psychiatry, 53*(1), 97-105.

Kratochwill, T. R., & Van Someren, K. R. (1985). Barriers to treatment success in behavioral consultation: Current limitations and future directions. *Journal of School Psychology, 23*(3), 225-239.

Krippendorff, K. (1980). *Content analysis: An introduction to its methodology.* Sage Publications: Beverly Hills, CA.

Lovaas, O. I. (1987). Behavioral treatment and normal educational and intellectual functioning in young autistic children. *Journal of Consulting and Clinical Psychology, 55,* 3–9.

Lecavalier, L., Leone, S., & Wiltz, J. (2006) The impact of behaviour problems on caregiver stress in young people with autism spectrum disorders. *Journal of Intellectual Disability Research, 50*(3), 172-183.

Matson, J. L., & Jang, J. (2014). Treating aggression in persons with autism spectrum disorders: A review. *Research in Developmental Disabilities, 35,* 3386-3391.

Matson, J. L., & Vollmer, T. (1995). *Questions about behavioral function (QABF).* Baton Rouge, LA: Disability Consultants, LLC.

Mazurek, M. O., Kanne, S. M., & Wodka, E. L. (2013). Physical aggression in children and adolescent with autism spectrum disorders. *Research in Autism Spectrum Disorders, 7,* 455-465.

McConachie, H., & Diggle, T. (2007). Parent implemented early intervention for young children with autism spectrum disorder: a systematic review. *Journal of Evaluation in Clinical Practice, 13*(1), 120-129.

McIntyre, L. L., Gresham, F. M., DiGennaro, F. D., & Reed, D. D. (2007). Treatment integrity of school-based interventions with children in the *Journal of Applied Behavior Analysis* 1991–2005. *Journal of Applied Behavior Analysis, 40,* 659–672.

Michie, S., Johnston, M., Francis, J., Hardeman, W., & Eccles, M. (2008). From theory to intervention: mapping theoretically derived behavioural determinants to behaviour change techniques. *Applied Psychology*, *57*(4), 660-680.

Michie, S., Richardson, M., Johnston, M., Abraham, C., Francis, J., Hardeman, W., ... & Wood, C. E. (2013). The behavior change technique taxonomy (v1) of 93 hierarchically clustered techniques: building an international consensus for the reporting of behavior change interventions. *Annals of Behavioral Medicine*, *46*(1), 81-95.

Najdowski, A. C., Wallace, M. D., Ellsworth, C. L., MacAleese, A. N., & Cleveland, J. M. (2008). Functional analyses and treatment of precursor behaviour. *Journal of Applied Behavior Analysis, 41*, 97-105.

Neidert, P. L., Dozier, C. L., Iwata, B. A., & Hafen, M. (2010). Behavior analysis in intellectual and developmental disabilities. *Psychological Services, 7*(2), 103-113.

Nevels, R. M., Dehon, E. E., Alexander, K., & Gontkovsky, S. T. (2010). Psychopharmacology of aggression in children and adolescents with primary neuropsychiatric disorders: a review of current and potentially promising treatment options. *Experimental and Clinical Psychopharmacology, 18*(2), 184-201.

Northup, J., Fisher, W., Kahng, S. W., Harrell, R., & Kurtz, P. (1997). An assessment of the necessary strength of behavioral treatments for severe behavior problems. *Journal of Developmental and Physical Disabilities, 9,* 1–16.

Parikh, M. S., Kolevzon, A., & Hollander, E. (2008). Psychopharmacology of aggression in children and adolescents with autism: a critical review of efficacy and tolerability. *Journal of Child and Adolescent Psychopharmacology, 18*(2), 157-178.

Petry, K., Maes, B., & Vlaskamp, C. (2005). Domains of quality of life of people with profound multiple disabilities: the perspective of parents and direct support staff. *Journal of Applied Research in Intellectual Disabilities, 18,* 35-46.

Reichow, B., Servili, C., Yasamy, M. T., Barbui, C., & Saxena, S. (2013). Non-specialist psychosocial interventions for children and adolescents with intellectual disability or lower-functioning autism spectrum disorders: a systematic review. *Plos Medicine*, *12*, 1-27.

Ryan, A. M., Leong, F. L., & Oswald, F. L. (2012). *Conducting multinational research: applying organizational psychology in the workplace.* Washington, D.C.: American Psychological Association.

Sallows, G. O., & Graupner, T. D. (2005). Intensive Behavioral Treatment for Children with Autism: Four-Year Outcome and Predictors. *American Journal on Mental Retardation*, *110*(6), 417-438.

Sanetti, L. M. H., & Kratochwill, T. R. (2007). Treatment integrity in behavioral consultation: Measurement, promotion, and outcomes. *International Journal of Behavioral Consultation and Therapy*, *4*(1), 95-114.

Schulte, A. C. (2007). Measurement in school consultation research. In W. P. Erchul & S. M. Sheridan (Eds.), *Handbook of research in school consultation: Empirical foundations for the field* (pp. 29-57). Mahwah, NJ:Erlbaum.

Sheinkopf, S. J., & Siegel, B. (1998). Home-based behavioral treatment of young children with autism. *Journal of Autism and Developmental Disorders, 28*(1), 15-23.

Shriver, M. D., Anderson, C. M., & Proctor, B. (2001). Evaluating the validity of functional behavior assessment. *School Psychology Review*, *30*(2), 180-192.

Sikora, D., Moran, E., Orlich, F., Hall, T. A., Kovacs, E. A., Delahaye, J., & ... Kuhlthau, K. (2013). The relationship between family functioning and behavior problems in children with autism spectrum disorders. *Research in Autism Spectrum Disorders*, *7*, 307-315.

Sim, J., & Wright, C. C. (2005). The kappa statistic in reliability studies: use, interpretation, and sample size requirements. *Physical Therapy*, *85*(3), 257-268.

Sim, J., & Reid, N. (1999). Statistical inference by confidence intervals: Issues of interpretation and utilization. *Physical Therapy*, *79*(2), 186-195.

Smith, R.G. (2011). Developing antecedent interventions for problem behavior. In W.W. Fisher, C. Piazza, & H. Roane (Eds), *Handbook of Applied Behavior Analysis* (pp. 297-316). NY: Guilford Publications.

St. Peter Pipkin, C., Vollmer, T. R., & Sloman, K. N. (2010). Effects of treatment integrity failures during differential reinforcement of alternative behavior: A translational model. *Journal of Applied Behavior Analysis, 43,* 47–70.

Strauss, K., Mancini, F., Fava, L., & SPC Group. (2013). Parent inclusion in early intensive behavior interventions for young children with ASD: a synthesis of meta-analyses from 2009 to 2011. *Research in developmental disabilities*, *34*(9), 2967-2985.

Sturmey, P. (2012) Treatment of psychopathology in people with intellectual and other disabilities. *Canadian Journal of Psychiatry, 57*(10), 593-600.

Sullivan, W. F., Berg, J. M., Bradley, E., Cheetham, T., Denton, R., Heng, J., … McMillan, S. (2011). Primary care of adults with developmental disabilities: Canadian consensus guidelines. *Canadian Family Physician*, *57*(5), 541–553.

Tager-Flusberg, H. (2014). Autism spectrum disorder: Developmental approaches from infancy through early childhood. In M. Lewis & K. D. Rudolph (Eds.), *Handbook of Developmental Psychopathology* (pp. 651-664). NY: Springer.

Thomason-Sassi, J. L., Iwata, B. A., Neidert, P. L., & Roscoe, E. M. (2011). Response latency as an index of response strength during functional analyses of problem behaviour. *Journal of Applied Behavior Analysis, 44,* 51-67.

Thompson, R. H., & Iwata, B. A. (2007). A comparison of outcomes from descriptive and functional analyses of problem behavior. *Journal of Applied Behavior Analysis*, *40*(2), 333–338. http://doi.org/10.1901/jaba.2007.56-06.

Weeden, M., Ehrhardt, K., & Poling, A. (2009). Conspicuous by the their absence: studies comparing and combining risperidone and applied behavior analysis to reduce challenging behavior in children with autism. *Research in Autism Spectrum Disorders, 3,* 905-912.

Wheeler, J. J., Baggett, B. A., Fox, J., & Blevins, L. (2006). Treatment integrity: a review of intervention studies conducted with children with autism. *Focus on Autism and other Developmental Disabilities*, *21*(1), 45-54.

Wightman, J., Julio, F., & Virués-Ortega, J. (2014). Advances in the indirect, descriptive, and experimental approaches to the functional analysis of problem behavior. *Psicothema, 26*(2), 186-192.

Wilder, D. A., Atwell, J., & Wine, B. (2006). The effects of varying levels of treatment integrity on child compliance during treatment with a three-step prompting procedure. *Journal of Applied Behavior Analysis,39*(3), 369-73.

Zaidman-Zait, A. a., Mirenda, P., Duku, E., Szatmari, P., Georgiades, S., Volden, J., & ... Thompson, A. (2014). Examination of bidirectional relationships between parent stress and two types of problem behavior in children with autism spectrum disorder. *Journal of Autism & Developmental Disorders, 44*(8), 1908-1917.

Appendix A

1

BEHAVIOURAL ASSESSMENT AND TREATMENT TAXONOMY (BATT)

Client ID:_____ D.O.B: _____ Gender:_____

Target Behaviour	Graphed? ✓ ×
Appropriate (to increase):	
Inappropriate (to decrease):	

Please review each item on the measure and indicate if it was utilized in the file you are reviewing.

ASSESSMENT CATEGORY

(1) NECESSARY ASSESSMENTS REVIEWED:	Documented ✓ ×
Medical "FSA-cited report by Dr B in May 2012"	
Psychiatric	
Developmental	
Academic	
Cognitive	
Communication	
Occupational Therapy	
Sensory Assessment	
Other:	
Current School Status	
Current Medical Status	

(2) MEDICATION USE

Name of Medication	Purpose	Dosage	Prescribed By?	Start Date DD/MM/YYYY	Discontinued DD/MM/YYYY

(3) SKILLS ASSESSMENTS:	Used ✓ × n/a
Adaptive Behaviour (e.g., VINELAND II, SIB-R, BASC):	
Children Sleep Habits Questionnaire	
Communication (e.g., VB MAPP, PPVT, EVT, CELF):	
Daily Living Skills (e.g., ABLLS):	
Functional Communication Record	
Indirect preference assessment with caregiver/child	
Screening Tool for Feeding Problems	
Sexuality (e.g., Boundaries):	
Skill Specific (e.g., Toileting):	
Social Skills (e.g., SSRS):	
Social Skills Rating System (Parent/Student - not entre section)	
Toileting Training readiness Checklist	
VB-MAPP (Bathing, Grooming, Dressing - not entire section)	
Other:	

(4) BEHAVIOURAL / PSYCHIATRIC MEASURES	Used ✓ × n/a
Dual Diagnosis Screen (e.g., Reiss, ADD)	
Severity Scale of Problem Behaviour (e.g, BPI, Aberrant Beh Checklist)	
Specific Diagnostic Screen (e.g., Y-BOCS, Conner's, BDI)	
QOL	
Other:	

(5) FUNCTIONAL BEHAVIOURAL ASSESSMENTS	Used ✓ × n/a
Indirect Measures	
Interviews	
Behavioural Contextual Interview (semi-structured)	
Bio-Psycho-Social Interview (semi-structured)	
Other:	

Questionnaires	
Questions About Behavioural Function (QABF)	
FAST	
MAS	
Target Behaviour Rating Scale (FIDD)	
Other:	
Direct Measures	
ABC Incident Recording BY Consultant/Tech	
ABC Incident Recording BY Caregiver/Mediator	
ABC Incident Recording Unspecified	
Functional Assessment Observation Form	
Informal in-vivo probe (i.e., *"Behaviour consultant asked caregiver to_____"*)	
Other:	
Experimental Methods (e*vidence of written description of conditions required*)	
Functional Analogue Assessment (antecedent & consequence manipulations) - FREQUENCY	
Functional Analogue Assessment (antecedent & consequence manipulations) - LATENCY	
Structural Analogue Assessment (antecedent manipulations)	
Other:	

(6) FBA RESULTS

Specific Measure Used (e.g., QABF)	**Escape**	**Attention**	**Tangible**	**Sensory**	**Physical**	**Automatic**		**Social**	
						R+	**R-**	**R+**	**R-**
Overall Hypothesized Function: (1 = Primary, 2 = Secondary)									

4

TREATMENT CATEGORY
(1) ANTECEDENT STRATEGIES:

(2) TEACHING ADAPTIVE BEHAVIOURS / SKILL BUILDING:	
SKILL (e.g., requesting _____, toilet training, etc.)	**METHOD** (e.g., Chaining, Shaping, Prompt Hierarchy, etc.)

5

(3) CONCEQUENCE STRATEGIES:

(4) MEDIATOR TRAINING	Used ✓ × n/a
Behaviour Skill Training (BST) Model	
Instruction/Reviewed Written Program	
Modeling – Live	
Modeling - Video	
Role-play / Practice	
Feedback	
Other:	
Adherence Measure _____% Adherence after training	

6

DATA COLLECTION CATEGORY
MONITORING OF TREATMENT: (Recording system used after treatment introduced)

CASE SUMMARY SECTION					
COMPLETE FOR BEHAVIOUR TARGETTED FOR INCREASE/DECREASE					
GOAL:					
DIRECTION OF DESIRED EFFECT (circle):	INCREASE		DECREASE		
TYPE OF DATA COLLECTED: (e.g. Frequency, duration including interval such as seconds, minutes etc, interval, partial interval, rating scale, percentage occurrence [include dimension i.e. frequency, duration etc.])					
TYPE OF DATA GRAPHED: (i.e. Daily, Weekly)					
BEHAVIOUR 1: _____					
UP TO 5 **BASELINE** DATA POINTS (LAST 5):					
UP TO 5 **TREATMENT** DATA POINTS (LAST 5):					
UP TO 5 **FOLLOW UP** DATA POINTS (LAST 5):					
BEHAVIOUR 2: _____ (*if applicable*)					
UP TO 5 **BASELINE** DATA POINTS (LAST 5):					
UP TO 5 **TREATMENT** DATA POINTS (LAST 5):					
UP TO 5 **FOLLOW UP** DATA POINTS (LAST 5):					